Innov8 Experiment Profile

Experiment Name	**Voice Analytics**	**Experiment Reference**	
Version	**1.0**	**Date**	**29 March 2010**
Experiment Owner	[redacted]	**Department**	**T12**
Experiment Start Date	**5 July 2010**	**Experiment End Date**	**August 2010**

Part 1 – Experiment Overview

Business Case

Detail the nature of the Experiment as a high level plain English description. Refrain from using overly technical terms if possible. Justify briefly why this Experiment should run, with a high level summary of Experiment, Conditions, De-risking effort, Business Case and Expected Outcomes/ Benefits.

As part of the Joint Capability Activity (JCA) in partnership with GCHQ, experiments by Voice Analytics are to be conducted within the Joint Collaboration Environment (JCE) using DISTILLERY. VoiceRT is NSA's currently deployed voice analytics technology that delivers integrated speaker ID (SID), language ID (LID), gender ID (GID), dual-tone multi-frequency (DTMF) detection, speech activity detection (SAD), and phonetic keyword search (i.e., direct keyword search on audio data) in more than 25 key foreign languages. VoiceRT requires specific hardware, proprietary process provisioning software (Tibco), and uses an older voice analytics technology. A new initiative is underway to create the next generation of VoiceRT that operates on commodity hardware (GHOSTMACHINE), replaces the Tibco process provisioning software with DISTILLERY, and makes use of the latest voice analytics technology from R64 (including the new Speech-to-Text). Voice Analytics experiments will occur in phases with this proposal being the first phase.

The business case for running Phase 1 of the Voice Analytics experiment in the JCE is to demonstrate the following:

- LPT voice data can be ingested into GHOSTMACHINE and passed to DISTILLERY for processing at the rate of arrival.
- DISTILLERY can run a subset of voice analytics (SAD, GID, and LID) on voice data.
- Running multiple analytics on voice data in parallel produces the same result as running each analytic independently.
- DISTILLERY can pass results from an analytic as input to a follow-on analytic with decision points throughout the dataflow based on analytic results (e.g. If GID reports the speaker is male, pass the data to LID; otherwise, stop processing the voice cut).
- DISTILLERY is capable of resource management at high speeds. If a voice analytic process in DISTILLERY is unable to handle the ingest rate, DISTILLERY can spawn additional voice analytic processes. Similarly, if resources from one set of voice analytic processes can be better used elsewhere, DISTILLERY will remove from memory some/all of the set of voice analytic processes.
- Determine the scalability as additional analytics are loaded into memory and processing data. How is the previous bullet affected as free memory goes to zero? How quickly can voice analytic processes be spawned and broken down?

Benefits

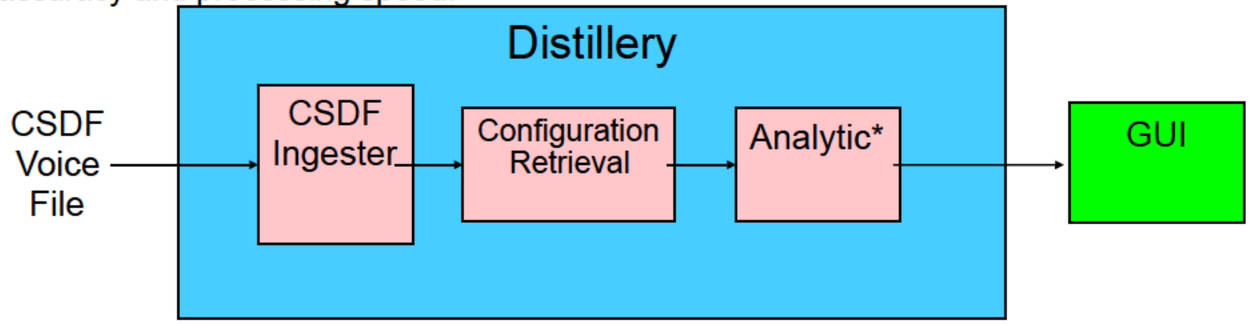
Outcome	Expected Benefit	How will benefit be measured?	KPI (CSF)	Reference Number
Demonstrate the JCE environment (including GHOSTMACHINE and DISTILLERY) can get data to core analytics at ingest rate.	The JCE environment can be used to process voice content. GHOSTMACHINE and DISTILLERY can maintain speeds required to processes voice on a 10G link. Add DISTILLERY to GHOSTMACHINE stack.	Monitor ingest pick-up directory that feeds DISTILLERY input.		
Demonstrate the JCE architecture can affordably scale in line with other VoIP accesses.	If the outcome is true, then the JCE environment provides a good test environment for future voice processing experiments. If VoiceRT is to eventually be replaced with this new technology, then tests must be run against a similar environment. This experiment also tests DISTILLERY's ability to dynamically manage resources, which is a requirement for our voice processing.	Compare quantity of VoIP data processed at JCE against that collected on similar links at NSAW and GCHQ.		
Demonstrate voice analytic processing of incoming data significantly reduces the amount of unintended collection reaching long-term stores by filtering on Language ID, Speech Activity Detection, and Gender ID.	Analysts can be directed to smaller, more probable collected voice cuts. Since voice files can be very large, significantly reduces the amount of long-term storage required.	DISTILLERY will be configured to separate wanted and unwanted VoIP files in separate storage locations. Comparing sizes of the two locations will provide an approximation for the ratio of wanted versus unwanted traffic.		

Demonstrate voice processing can be performed on commodity hardware (GHOSTMACHINE). Current voice processing at NSA has hardware specific requirements.	Voice analytics can be executed on commodity hardware alongside other applications. Allows for easier hardware maintenance and shorter deployments.	Analysts will confirm resulting data has been filtered without a loss of wanted traffic.		
JCE experiments can process content while remaining in compliance with US and UK laws.	Future experiments requiring content for processing can leverage.	If MONKEYPUZZLE can safely produced selected content to the experiment, then it should be in compliance.		

Plan
Timeline, including dependencies/pre-requisites, set-up, operation, evaluation, reporting, closedown. Also includes deliberate outages of equipment.

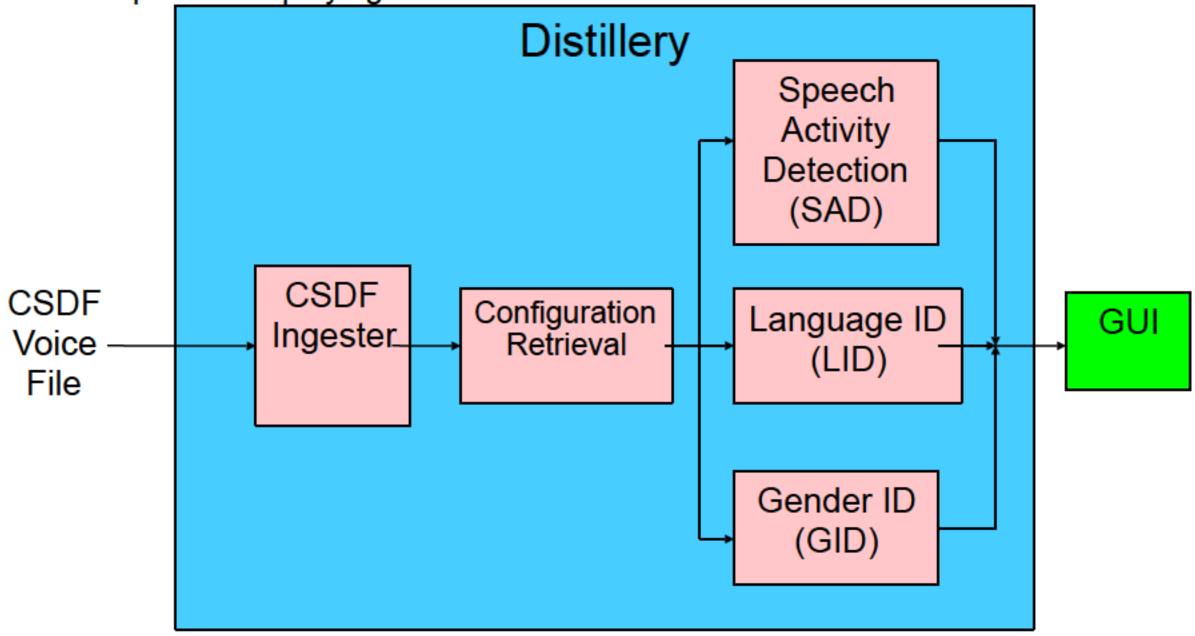
As stated in the business case, this is Phase 1 of the Voice Analytics Experiment. This phase is divided into four sub-experiments focusing on dataflow through a select group of voice analytics (SAD, GID, LID) in DISTILLERY. All four experiments require access to CSDF voice files. Future phases will consider SOTF once GHOSTMACHINE can support this format.

Exp. 1: Independent Analytic Tests (~1 week) – Designed to test the dataflow from DISTILLERY input, through a few processing blocks, and the output displayed to a GUI. Each analytic is executed independently as a single instance to determine accuracy and processing speed.

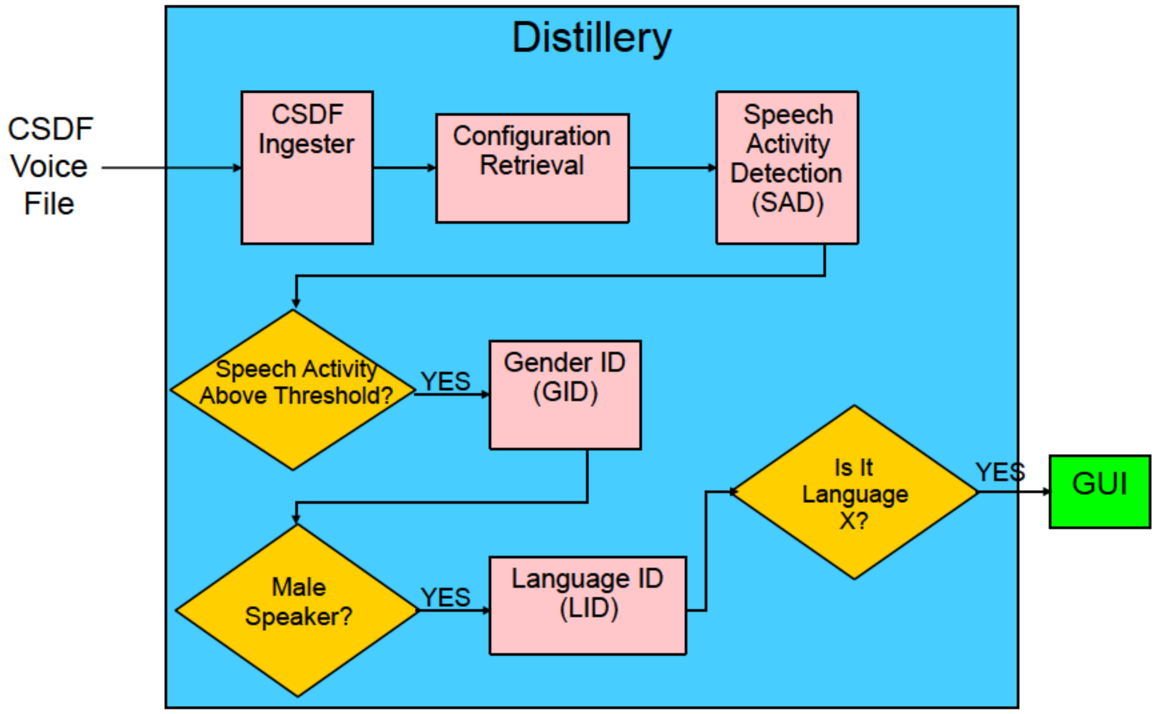


* Speech Activity Detection (SAD), Language ID (LID), and Gender ID (GID) will each be tested independently as the active analytic

Exp. 2: Parallel Analytic Tests (~1 week) – Similar to first experiment, but each CSDF voice file is run against each of the three voice analytics simultaneously. This should confirm data is not corrupted by multiple processes accessing simultaneously. Experiment also adds the complexity of holding results until all analytics have executed prior to displaying on GUI.

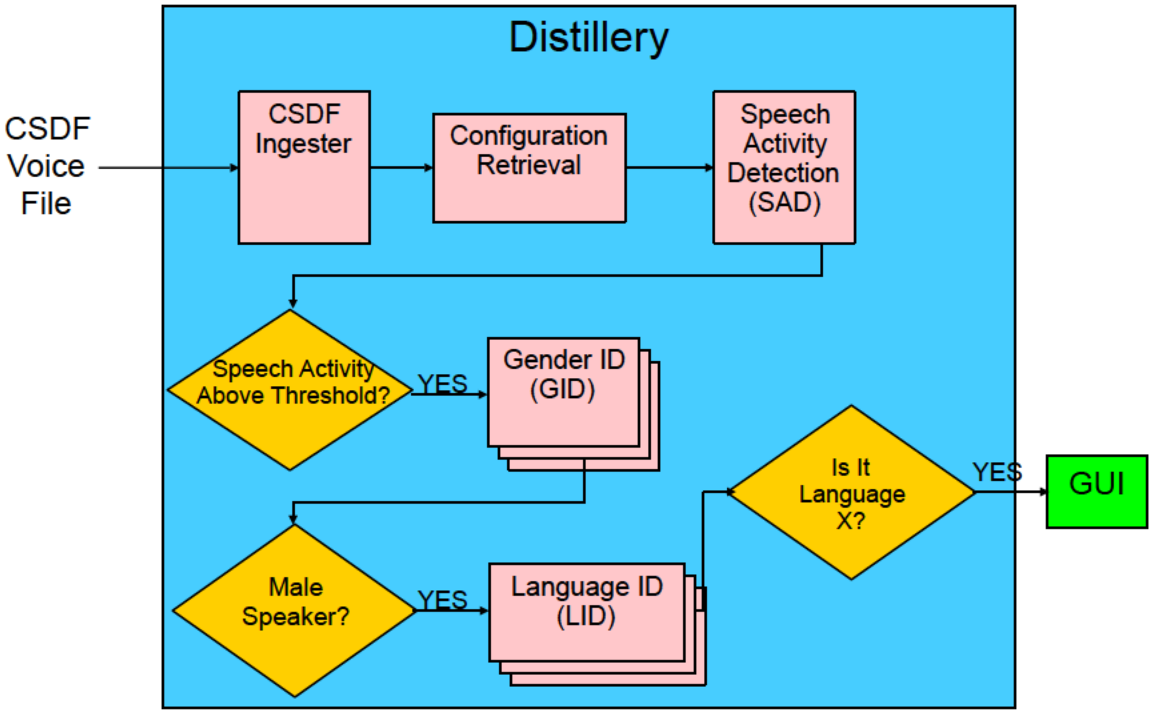


Exp. 3: Linked Analytic Tests (~1 week) – Similar to the previous experiments, but results from the voice analytics are now used to make decisions about future processing of each file. The ingest rate should be fast enough to create a noticeable bottleneck in flow.



This information is exempt from disclosure under the Freedom of Information Act 2000 and may be subject to exemption under other UK information legislation. Refer disclosure requests to GCHQ o [redacted] (non-sec) or emai [redacted] chq

Exp. 4: Dynamic Resource Manager Tests (~2 week) – Builds off of the third experiment. The SAD algorithm is much faster than GID and LID. In this experiment, DISTILLERY should be able to dynamically create/remove additional GID and LID analytic processes to compensate for GID and LID processes falling behind or becoming idle. A more advanced part of this experiment would be to monitor DISTILLERY's response to all system memory in use; for example, 100 GID analytics and 150 GID analytics are still falling behind, but no system memory exists to spawn additional processes. Does one set of analytic processes starve or do we see a constant exchange of memory between GID and LID?


Distillery
CSDF Voice File
CSDF Ingester
Configuration Retrieval
Speech Activity Detection (SAD)
Speech Activity Above Threshold?
YES
Gender ID (GID)
Is It Language X?
YES
GUI
Male Speaker?
YES
Language ID (LID)